Modellgetriebene semi-automatische Bereitstellung von Analytischen Informationssystemen

Yvette Teiken

30. April 2009

Zusammenfassung

Analytische Informationssysteme werden von Organisationen immer häufiger zur Informationsversorgung und Entscheidungsunterstützung eingesetzt [Fri07]. Vielfach werden hierfür Analytische Informationssysteme mit zu Grunde liegender Informationslogistik (IL)-Infrastruktur verwendet. IL-Infrastruktur beschreibt den bei Analytischen Informationssystemen über Insellösungen in Fachabteilungen hinaus gehenden Teil. Hierzu gehören ein integrierter Datenbestand sowie eine Analyse-Sicht z.B. in Form von OLAP-Servern. Notwendige Schritte zur Bereitstellung von IL-Infrastruktur sind mit einem hohen Aufwand verbunden und somit kostenintensiv. In diesem Artikel wird ein Ansatz zur einfacheren Bereitstellung von IL-Infrastruktur vorgestellt. Dies soll erreicht werden, indem Teile generativ erzeugt werden. Hierbei wird eine bedarfsgetriebene Analyse vorausgesetzt. Realisiert werden soll ein Top-Down-Vorgehen, das mittels Domänen-spezifischer Modellierung (DSM) erreicht wird. In diesem Vorgehen sollen Aspekte von IL-Infrastruktur mittels einer Familie von Domänenspezifischen Sprachen (DSL) beschrieben werden. Durch dieses Vorgehen können Experten der jeweiligen Domäne Teile der IL-Infrastruktur entwerfen und erzeugen lassen. Generiert werden u.a. Kennzahlen, multidimensionale Sicht, Ladeprozesse, Sicherheit und auch dynamische Berichte.

1 Motivation/Einführung

Organisationen stehen oftmals unter ständig steigendem Kostendruck. Um diesem Kostendruck gerecht zu werden, spielt ein effektives Controlling eine immer wichtigere Rolle. Zunehmend wird dabei auf IT-Unterstützung gesetzt. Diese Art von Systemen werden als Analytische Informationssysteme bezeichnet. Die Analytischen IS werden nach [CG04] in verschiedene Reifegrade unterteilt, die sich in ihrer Integrationshöhe unterscheiden. Für eine Organisation ist der Übergang von fachspezifischen Lösungen hin zu organisationsweiten Lösungen mit hohem Aufwand und somit hohen Kosten verbunden. Hier muss die Integration verschiedener Fachbereiche vorgenommen werden. Dazu gehört der Aufbau eines übergreifenden Metadatenmanagements, die Integration externer Daten und die Definition von Analyseszenarien. Nach [BG05] sind zwei verschiedene Vorgehen zu unterscheiden, das Informationsbedarf-getriebene und das Daten-getriebene Vorgehen ([Muc06]). Bei einem Informationsbedarf-getriebenen Vorgehen wird die Entwicklung des Analytischen Informationssystems auf Basis der inhaltlichen Anforderungen spezifiziert. Diese Art des Vorgehens wird als Top-Down-Vorgehen bezeichnet. Beim Top-Down-Vorgehen stellt die Analyse des Informationsbedarf ein Kernaufgabe dar.

In diesem Beitrag wird ein Top-Down-Vorgehen für die modellgetriebene semi-automatische Bereitstellung von Analytischen Informationsystemen vorgestellt. Hierfür wird im nächsten Kapitel erst die Idee und dann der Lösungsansatz beschrieben.

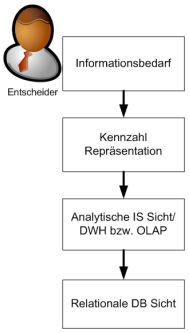

Abbildung 1: Sichten im TopDown Vorgehen

2 Lösungsansatz

Um aus Informationsbedarf eine IL-Infrastruktur eines Analystischen IS abzuleiten zu können, werden verschiedene Zwischenschritte absolviert. Diese Schritte sind schematisch in der Abbildung 1 abgebildet. Ein Informationsbedarf in einem Analytischen IS kann mittels relevanter Kennzahlen beschrieben werden. Dieser Informationsbedarf wird von einem Entscheider formuliert. Dieser richtet sich nach den aktuellen Gegebenheiten der Organisation. Gegebenenfalls werden relevante Grenzen oder Schwellwerte formuliert. Bei einer Kennzahl kann es sich um einfache Kennzahlen handeln, die sich zum Beispiel mittels einer Division beschreiben lassen. Dies ist bei der betrieblichen Kennzahl Umsatzrendite der Fall ($Umsatzrendite = \frac{Gewinn}{Umsatz}$). Ein Informationsbedarf kann jedoch auch durch eine komplexe Kennzahl beschreiben werden. Zu diesen komplexen Kennzahlen gehören epideminologische Kennzahlen, wie sie im Rahmen der Arbeit des EKN[1] verwendet werden. Hier wird mit Erwartungswerten gearbeitet, die Konstrukte wie eine Standardbevölkerung voraussetzen, wie zum Beispiel *Erwartete Fallzahlen* ($e = \sum_{a \in A} R_a * n_a$). Diese Kennzahl stellt eine Art Prognose der zu erwartenden Fallzahlen dar, dieser werden auf Basis einer Bevölkerungsprognose, Fallzahlen und einer externen Bevölkerung ermittelt.

Kennzahlen lassen sich formalisieren. Auf Grundlage der formalisierten Kennzahlen ist eine Formalisierung des Informationsbedarfs möglich. Werden diese formalisierten Kennzahlen mit Dimensionsinformationen angereichert, so lässt sich eine Analytische IS-Sicht in Form einer DWH Sicht erzeugen. Hierbei entsteht implizit ein übergreifendes Metadatenmanagement.

Aus der Analytischen Sicht lassen sich danach die relationalen Strukturen in Form von Datenbank-Sichten erzeugen.

2.1 Konkreter Lösungsansatz

In dem in Abbildung 2 skizzierten Ansatz lassen sich unterschiedliche Rollen bzw. Teilaufgaben identifizieren. So werden Kennzahlen zum Beispiel von Entscheidern modelliert, während Metadaten von Modellierungsexperten beschrieben werden. Aus diesem Grunde erscheint es sinnvoll, die Bereitstellung von Analytischen IS in verschiedene Teilaspekte zu zerlegen. Diese Teilaspekte können dann von den jeweiligen Experten/Rollen beschrieben werden, so dass am Ende eine integrierte Gesamtsicht entsteht. Die zu modellierenden Teilaspekte unterscheiden sich durch Vorwissen/Vorbedingungen und unterschiedlich hohes Abstraktionsniveau. Deswegen ist es nicht sinnvoll, diese Aspekte alle auf die gleiche Art zu beschreiben. Deshalb soll ein DSM basierter Ansatz für die Erzeugung von Analytischen IS verfolgt werden (siehe auch [KT08]). Hierfür soll eine Familie von grafischen und textuellen DSLs entwickelt werden, ähnlich wie es in [WK06] durchgeführt worden ist. Die Verwendung von DSLs hat den Vorteil, dass Modelle von Fachexperten in ihrer eigenen Begriffswelt erstellt werden können und keine zusätzlichen Sprachen vom Fachexperten für die DWH-Konzeption, wie zum Beispiel UML, erlernt werden müssen. UML

[1] http://www.krebsregister-niedersachsen.de/

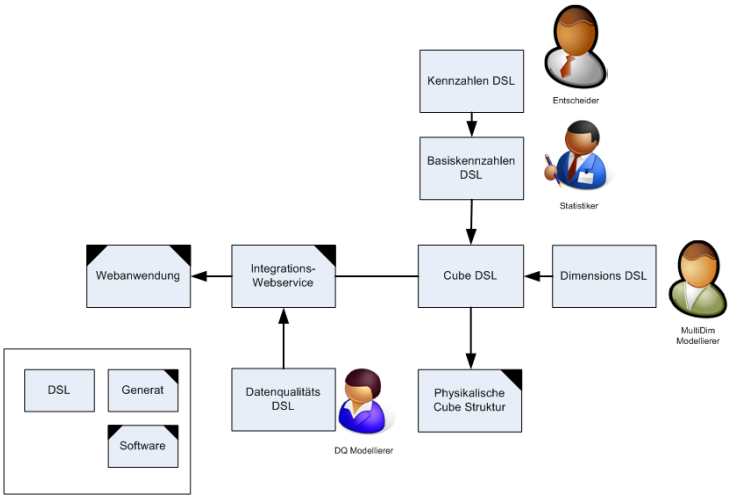

Abbildung 2: Familie von DSLs und deren Generate

ist zwar ebenfalls eine DSL, allerdings für den Einsatzbereich Softwareentwicklung, und somit nur bedingt für DWH-Fachexperten geeignet. Beispiele für Modellierungssprachen im Bereich konzeptueller Modellierung stellen ADAPT [Bul96] und ME/R [SBHD99] dar. Sie dienen als Grundlage der späteren grafischen DSL.

In der Abbildung 2 sind die mögliche Familie von DSLs und deren Generate abgebildet. Die Kennzahlen DSL dient zur Beschreibung von Kennzahlen und deren Verknüpfung miteinander. Diese DSL soll möglichst einfach gehalten sein und sich rein auf die Kennzahl-Verknüpfung beschränken. Die Basiskennzahlen DSL hingegen ist komplexer. Sie beschreibt die multidimensionale Annotation von Kennzahlen. Ziel der Kennzahlmodellierung ist die Überführung in eine multidiemsionale Struktur. Je nach Art der Organisation und gegebenen Umständen kann eine Kennzahl unterschiedliche Dimensionen aufweisen. Zum Beispiel kann die Kennzahl *Gewinn* bei einer Supermarktkette andere Ausprägungen haben als bei einem Krankenhaus. Eine Supermarktkette wird ihren Gewinn nach OPS- oder ICD-Dimensionen annotieren. Es kann auch vorkommen, dass die gleiche Dimension verwendet wird, diese jedoch unterschiedliche Bedeutungen hat. So kann eine geografische Dimension bei einer Supermarktkette das Fillialnetz beschreiben während bei einem Krankenhaus das Einzugsgebiet gemeint ist. In dem hier vorgestellten Ansatz wird diese Art von Semantik nicht berücksichtigt, da beide Bedeutungen das gleiche schematische Konzept beschreiben.

Die Cube DSL beschreibt die Kennzahl aus multidimensionaler Sicht. Es wird die multidimensionale Modellierungssprache ADAPT verwendet. Sie dient dazu, eine einfache Überführung in eine multidimensionale Sicht zu ermöglichen. Weiterhin wird hier die mögliche Verknüpfung von Cubes auf Zellebene beschrieben. Dies wird benötigt, um die Operationen, die mit den Kennzahlen beschrieben sind, auf Cube Ebene zu realisieren. Dies ist vom Prinzip ähnlich, wie das in [GKS09] vorgeschlagene Vorgehen.

Für die Beschreibung von Dimensionen existiert ebenfalls eine DSL. Zwar besteht in ADAPT und anderen Modellierungssprachen die Möglichkeit, Dimensionen zu beschreiben, doch da die Modelle sehr komplex sind, empfiehlt sich die Verwendung einer eigenen DSL. In dieser DSL ist es möglich, komplexe Dimensionen mit parallelen Hierarchien mit vielen hunderten Knoten zu Elementen zu modellieren. Hierfür müssen besondere Formen der Eingabe und Darstellung beachtet werden. Die in der Dimensions DSL beschrieben Dimensionen werden in der Cube DSL referenziert.

Die alleinige Bereitstellung einer DSL oder einer Familie von DSLs stellt allerdings noch keinen hohen Mehrwert für die Realisierung eines Projekts dar. Dies geschieht erst dann, wenn die mit Hilfe einer DSL erstellten Modelle semi-automatisch zu neuen Modellen oder zu Generaten,

wie sie in Abbildung 2 dargestellt sind, transformiert werden. Dabei soll es auch möglich sein, aus einem Modell mehrere Artefakte für unterschiedliche Einsatzzwecke zu erzeugen. Dieser Ansatz wird in der Softwareentwicklung als MDSD bezeichnet (siehe [SVE07]). Dieser soll im Rahmen der Arbeit auf die Bereitstellung von Informationslogistik übertragen werden, um die oben genannten Ziele zu realisieren.

Sichtbare Elemente einer DSL ist die konkrete Syntax. Die eigentliche Bedeutung bzw. Interpretation wird in der abstrakten Syntax der DSL spezifiziert. Diese abstrakte Syntax wird mittels eines Metamodells beschrieben. Werden, wie in diesem Ansatz vorgestellt, mehrere DSLs realisiert, die miteinander interagieren so stellt sich die Frage ob ein gemeinsames oder getrennte Metamodelle verwendet werden sollen (siehe [HCW07]). Da es in dem in dieser Arbeit beschriebenen Ansatz viele überschneidende Elemente gibt, wird ein gemeinsames Metamodell verwendet. Dies bedeutet, dass Modell-zu-Modell-Transformationen Modell-inhärent sind. Mit Hilfe dieser Transformationen lassen sich andere Modelle oder Artefakte erzeugen. In Abbildung 2 sind neben der Familie der DSLs ebenfalls die möglichen Generate abgebildet.

Oftmals wird ein multidimensionales Schema in einer relationalen Datenbank abgelegt. Dies wird mit dem Generat der physikalischen Cube Struktur beschrieben. Für die Repräsentation eines multidimensionalen Schemas existieren verschiedene relationale Abbildungen wie zum Beispiel das Snow- oder Star-Schema (siehe [BG04]). Ein anderes Beispiel für eine solche Abbildung ist das MUSTANG Schema aus [KMR03], welches im OFFIS entwickelt worden ist. Dies ist besonders gut geeignet in Bezug auf epidemologische Fragestellungen. Dadurch, dass in diesem Ansatz ein MDSD-Ansatz verwendet wird, können durch das Anwenden anderer Transformationen auf Basis der Cube DSL andere relationale Modelle erzeugt werden. Es muss nur die Transformation von einem kleinen Teil des Metamodells aus spezifiziert werden.

Ist die Datenstruktur bekannt, die innerhalb der DSLs beschrieben worden ist, so ist auch bekannt, wie die Daten aussehen müssen, die für die Auswertungen benötigt werden. Somit ist es möglich, aus der Cube DSL einen Webservice zu erzeugen, der die spezifizierten Daten entgegen nimmt und diese im Analytischen IS ablegt. Dieser Service stellt dann sicher, dass die integrierten Daten korrekt sind. Allerdings bezieht sich Korrektheit nur auf Teilaspekte der Daten wie Datentyp, ungültige Werte oder Minimum und Maximum. Echte Datenqualität, wie zum Beispiel in [Brü08] beschrieben, kann auf diese Weise nicht erreicht werden. Dies wird mittels der Datenqualitäts DSL beschrieben. Weiterhin ist es denkbar dieses Konzept schrittweise zu erweitern, um somit Aspekte wie dynamische Berichte oder Ladeprozesse zu integrieren.

3 Zusammenfassung

Mit dem hier vorgestellten Ansatz lassen sich auf einfache Weise Analytische IS schneller und in höherer Qualität erzeugen. Weiterhin lässt sich durch diesen Ansatz ein Modell-Repository Aufbauen, welches den Modellierungsaufwand über die Zeit reduziert. Durch den erweiterbaren Ansatz können auch weiter gehende Konzepte wie beispielsweise Reporting integriert werden.

Literatur

[BG04] Bauer, Andreas ; Günzel, Holger: *Data-Warehouse-Systeme. Architektur, Entwicklung, Anwendung.* Dpunkt Verlag, 2004. – ISBN 3898642518

[BG05] Burmester, Lars ; Goeken, Matthias: Benutzerorientierter Entwurf von unternehmensweiten Data-Warehouse- Systemen. In: Ferstl, Otto K. (Hrsg.) ; Sinz, Elmar J. (Hrsg.) ; Eckert, Sven (Hrsg.) ; Isselhorst, Tilman (Hrsg.): *Wirtschaftsinformatik*, Physica-Verlag, 2005. – ISBN 3–7908–1574–8, S. 1421–1440

[Brü08] Brüggemann, Stefan: *Proaktives Management von Konsistenzbedingungen im Analytischen Performance Management.* 2008

[Bul96] Bulos, Dan: OLAP database design: A new dimension. In: *Database Programming&Design* Vol. 9(6) (1996)

[CG04] Chamoni, Peter ; Gluchowski, Peter: Integrationstrends bei Business-Intelligence-Systemen, Empirische Untersuchung auf Basis des Business Intelligence Maturity Model. In: *Wirtschaftsinformatik* 2 (2004), S. 119–128

[Fri07] Friedrich, Dirk: Einfach soll es sein - bei hoher Datenqualität. In: *IS - Informationsplattform für Business Applications* 11 (2007), S. 30–35

[GKS09] Gluchowski, Peter ; Kunze, Christian ; Schneider, Christian: A Modeling Tool for Multidimensional Data using the ADAPT Notation. In: *42nd Hawaii International Conference on System Sciences (HICSS-42)*, 2009

[HCW07] Hessellund, Anders ; Czarnecki, Krzysztof ; Wasowski, Andrzej: Guided Development with Multiple Domain-Specific Languages. In: Engels, Gregor (Hrsg.) ; Opdyke, Bill (Hrsg.) ; Schmidt, Douglas C. (Hrsg.) ; Weil, Frank (Hrsg.): *MoDELS* Bd. 4735, Springer, 2007 (Lecture Notes in Computer Science). – ISBN 978–3–540–75208–0, S. 46–60

[KMR03] Koch, Sascha ; Meister, Jürgen ; Rohde, Martin: MUSTANG – A Framework for Statistical Analyses of Multidimensional Data in Public Health. In: *Proceedings of the 17th International Conference Informatics for Environmental Protection.* Cottbus, September 2003, S. 635–642

[KT08] Kelly, Steven ; Tolvanen, Juha-Pekka: *Domain-Specific Modeling: Enabling Full Code Generation.* John Wiley & Sons, 2008

[Muc06] *Kapitel* Das Data Warehouse als Datenbasis analytischer Informationssysteme. In: Mucksch, Harry: *Analytische Informationssysteme. Business Intelligence-Technologien und -Anwendungen: Business Intelligence-Technologien und -Anwendungen.* Springer, Berlin, 2006, S. 129–142

[SBHD99] Sapia, Carsten ; Blaschka, Markus ; Höfling, Gabriele ; Dinter, Barbara: Extending the E/R Model for the Multidimensional Paradigm. In: *ER '98: Proceedings of the Workshops on Data Warehousing and Data Mining.* London, UK : Springer-Verlag, 1999. – ISBN 3–540–65690–1, S. 105–116

[SVE07] Stahl, Thomas ; Völter, Markus ; Efftinge, Sven: *Modellgetriebene Softwareentwicklung. Techniken, Engineering, Management.* Dpunkt Verlag, 2007. – ISBN 3898644480

[WK06] Warmer, J. B. ; Kleppe, A. G.: Building a Flexible Software Factory Using Partial Domain Specific Models. In: *Sixth OOPSLA Workshop on Domain-Specific Modeling (DSM'06), Portland, Oregon, USA.* Jyvaskyla : University of Jyvaskyla, October 2006. – ISBN 951–39–2631–1, S. 15–22